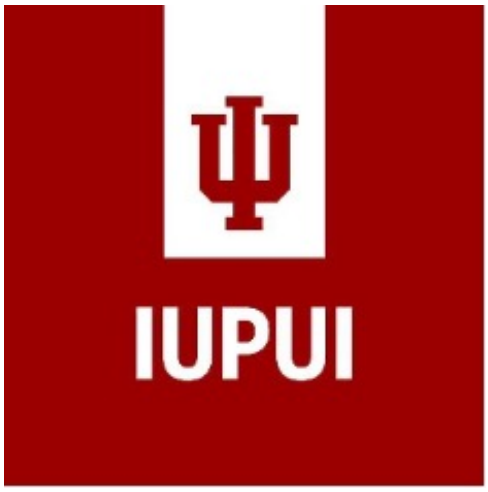
DermRAG: Fine-Tune Embedding model vs Naive Embedding Model

INFO-H 518 Deep Learning

By

Anusha Gadgil
Bhavya Kalra
Neeraj Gupta
Raj Acharya

Instructor
Dr. Sunandan Chakraborty



Indiana University - Purdue University Indianapolis
USA
May, 2024

Contents

1 Problem Statement

The goal is to develop an AI-based system called "DermRAG" that can effectively retrieve queries related to dermatological signs of systemic diseases. This is an important problem as accurately diagnosing skin conditions and their underlying systemic associations is crucial for effective treatment and patient well-being. However, navigating the vast amount of information available in medical literature can be challenging for healthcare professionals. DermRAG aims to bridge this gap by making knowledge more accessible and user-friendly through the use of large language models. While building the DermRAG, we discovered that there is a need for better retrieval to improve the generated answers. In our project, we are looking to improve the retrieval part of the RAG system by improving the embedding model used for embedding user queries and the indexed dataset.

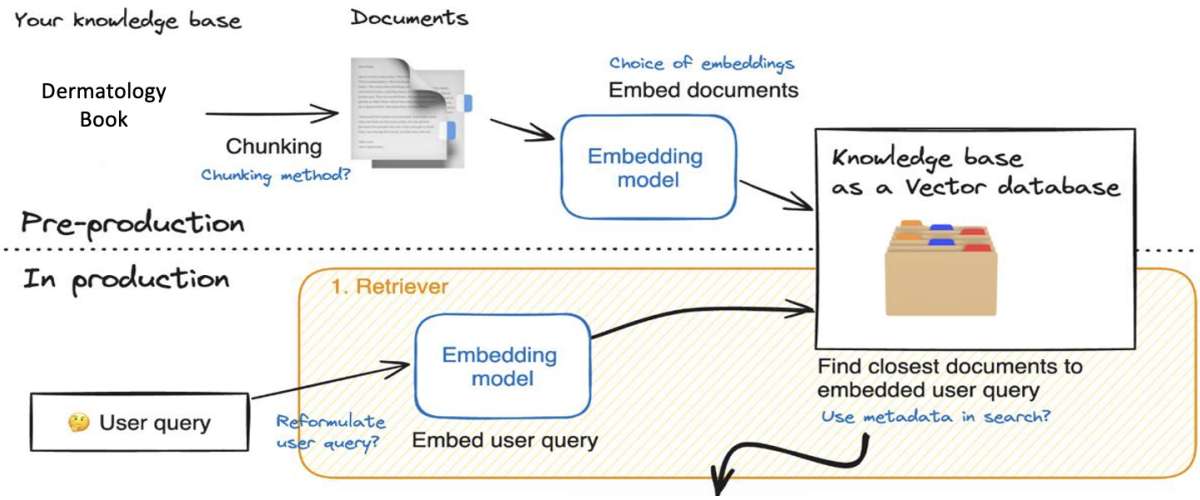


Figure 1: Retrieval module of RAG system

2 Dataset

The dataset consists of two key resources:

1. The book "Dermatological Signs of Systemic Disease, 5th Edition" by Callen et al. , which provides in-depth information on the cutaneous manifestations of systemic disorders. This authoritative text has become an essential reference for dermatologists, internists, and other healthcare professionals.

2. The synthetic dataset was created by dividing dataset(dermatology book) into 738 chunks. We generated 3 possible questions per chunk ChatGPT3.5-turbo model. This synthetic data is used to fine-tune the embedding model.

The dataset was constructed by carefully extracting relevant information from the book and formatting it into a structured Question-Chunk format.

3 Methodology

The process to fine-tune the embedding model is fragmented into the following structure.

1. **Data Collection and Extraction:** Extract relevant information from the book ”Dermatological Signs of Systemic Disease, 5th Edition” by Callen et al., including disease descriptions, symptoms, diagnostic criteria, and treatment recommendations.

2. **Data Preprocessing:** Chunk the dataset using a chunking process to divide it into chunks for further analysis. Utilize a SentenceSplitter to segment the text into individual sentences, facilitating easier processing and analysis.

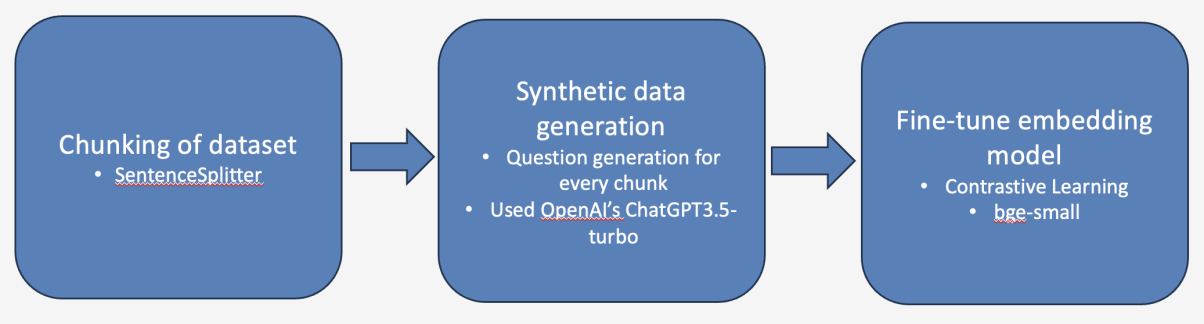


Figure 2: Overall fine-tuning process

3. **Synthetic Data Generation:** Generate synthetic questions for each chunk of text in the dataset using OpenAI's ChatGPT3.5-turbo. This step creates additional training examples to enhance the model's understanding and performance. Format the extracted data into a structured questionnaire format, creating the "Dermatology Question-Chunk Dataset: Skin Disease".

4. **Model Fine-tuning:** Fine-tune the embedding model (e.g., BGE-small) on the Dermatology Question-Chunk Dataset using the LlaMA-Index library. Split the dataset into training and validation sets using scikit-learn's train test split function with an 80-20 split ratio. Fine-tune the embedding model on the training set using the SentenceTransformersFinetuneEngine from LlaMA-Index. Leveraging **contrastive learning** objective during this process further augments the dataset by encouraging the model to learn from both positive and negative examples. By contrasting between relevant and irrelevant information, the model can better discern subtle distinctions and improve its ability to provide accurate responses to queries. This augmentation not only enriches the dataset but also enhances the model's generalization capabilities, leading to more robust performance across a variety of queries and scenarios.

5. **Model Evaluation:** Evaluate the fine-tuned model's performance on the val- idation set using two approaches:

   a. Custom Hit Rate Metric: Calculate the percentage of queries where the relevant document was retrieved among the top-k results.

   b. Information Retrieval Evaluator: Utilize the Information- RetrievalEvaluator from the Sentence Transformers library, which provides a comprehensive suite of metrics.

The fine-tuned model, named "DermRAG," will be saved for future use and evaluation.

4 Results

The results demonstrate the effectiveness of fine-tuning an embedding model on a synthetic (LLM-generated) dataset for an information retrieval task related to dermatological signs of systemic diseases. The evaluation is done in the following way,

1. Splitting the dataset into training and validation sets using scikit-learn's `train_test_split` function.

2. Generating synthetic questions for each text chunk in the corpus using GPT- 3.5-turbo. Each pair of (generated question, text chunk) was added to the fine-tuning dataset. It gives out **1773 queries for training and 447 for validation data. Giving 3 queries for each chunk.**

3. The dataset(dermatology book) is divided into **738 chunks** among training and validation data.

Table 1: Dataset Division

Data Split.	Chunks	Queries generated
80%	590	1773
20%	148	447

4. Fine-tuning the Embedding Model on the training set using the `SentenceTransformersFinetuneEngine` from LlaMA-Index.

After fine-tuning, the performance of three embedding models was evaluated: proprietary OpenAI embedding model for benchmarking; open-source BAAI/bge-small-en embedding model; fine-tuned embedding model. Embedding models are evaluated mainly on two evaluation criterions,

1. A simple hit rate metric that calculates the percentage of queries where the relevant document was retrieved among the top-k results.

2. The `InformationRetrievalEvaluator` from the sentence_transformers library, which provides comprehensive metrics for different top-k values.

The results indicate that the fine-tuned model outperforms others in various metrics:

Table 2: Model outcomes

Sr. No.	Model	Hit Rate	MAP
1	Open AI	98.31%	-
2	BAAI	94.59%	87.48%
3	Fine Tuned	95.60%	88.65%

Table 3: Model Accuracy

Model	Accuracy@1	Accuracy@3	Accuracy@5	Accuracy@10
Naive BGE/BAAI	81.08%	90.54%	94.59%	96.28%
Fine-tuned BGE/BAAI	81.41%	91.89%	95.60%	97.29%

Table 4: Model Precision

Model	Precision@1	Precision@3	Precision@5	Precision@10
Naive BGE/BAAI	81.08%	30.18%	18.91%	9.62%
Fine-tuned BGE/BAAI	81.41%	30.63%	19.12%	9.73%

5 Conclusion

In conclusion, the results obtained from fine-tuning the embedding model demonstrate significant improvements in retrieval performance within the RAG system, even when working with a relatively small dataset. The findings indicate that the fine-tuned model outperforms other models across various evaluation metrics.

Specifically, the fine-tuned model achieves a hit rate of 95.6% and a MAP score of 87.48%, showcasing its effectiveness in accurately retrieving relevant information. Moreover, the model consistently achieves higher Accuracy@1, Accuracy@3, Accuracy@5, and Accuracy@10 scores compared to alternative models, indicating superior retrieval of relevant documents in the top positions.

Additionally, the fine-tuned model demonstrates higher Precision@1, Recall@1, Precision@3, and Recall@3 scores, further emphasizing its ability to retrieve a larger proportion of relevant documents in the top positions. This is particularly crucial for enhancing the user experience and ensuring that the most relevant information is readily accessible.

Overall, these findings highlight the effectiveness of fine-tuning the embedding model for enhancing retrieval performance within the RAG system. Moving forward, leveraging larger datasets to further refine the model holds promise for achieving even greater improvements in retrieval accuracy and efficiency.

6 References

Literature review:

- Deep Learning in Dermatology: A Systematic Review of Current Approaches, Outcomes, and Limitations Link
- Hugging Face DocumentationLink
- LLaMA: Open and Efficient Foundation Language ModelsLink
- LLaMA-indexLink
- BAAI/bge-small-en Link
- OpenAI Link
- Dermatology Book: Jeffrey P. Callen, Joseph L. Jorizzo, John J. Zone, Warren Piette - Dermatological Signs of Systemic Disease (2016, Elsevier)Link

7 Appendix

Code development:

- Installed required packages for fine-tuning using LlaMA-Index library
- Defined a function to load corpus from files
- Split dataset into training and validation sets
- Generated synthetic questions using GPT-3.5-turbo for fine-tuning
- Fine-tuned Embedding Model on training set
- Evaluated performance of OpenAI, BAAI/bge-small-en, and fine-tuned models using hit rate metric and InformationRetrievalEvaluator

♣ Contributions

Table 5: Contributions

Work	Anusha Gadgil	Bhavya Kalra	Neeraj Gupta	Raj Acharya
Mid-Term literature review	✔	✔	✔	✔
Mid-Term coding and initial result		✔	✔	
Mid-Term PPT	✔	✔		✔
Final literature review	✔		✔	✔
Final coding and initial result	✔	✔	✔	
Final PPT and Report		✔	✔	✔